Predicting Heart Disease
Data 621: Data Mining Final Project

Betsy Rosalen, Gabrielle Bartomeo, Jeremy O’Brien, Lidiia Tronina, Rose Koh

CVD is leading cause of death globally

Cardiovascular disease (CVD) is responsible for **18MM deaths worldwide in 2015**

CVD includes heart attacks, strokes, heart failure, coronary artery disease, arrhythmia, venous thrombosis, and other conditions

47% of Americans have at least one of key risk factors: blood pressure. cholesterol, or smoking

Researchers estimate **90% of CVD deaths could be prevented**

More efficient, scalable, and non-invasive early detection can lead to medical interventions, preventive care, or behavioral change

Applying data mining techniques to predict risk based on existing or easy-to-collect health data could improve healthcare outcomes and mortality rates

CVD data mining is ongoing area of research

Shouman et. al conducted exhaustive review of classification work between 2000 and 2016

Wide range of classification techniques used on published CVD datasets:

- Logistic Regression
- Decision Trees
- Random Forests
- KNN
- Naïve Bayes
- Neural Networks
- Multilayer Perceptron
- Support Vector Machines
- Associative Classifiers

Diagnostic accuracy of classifier models built on the Cleveland dataset peaks in the .80 range:

Technique	Median Accuracy (n = 62 studies)
Logistic Regression	0.855
Random Forest	0.724
Support Vector Machine	0.809
Naive Bayes	0.819

Cleveland Heart Disease Dataset

The Cleveland dataset contains 303 observations and 76 attributes in total.

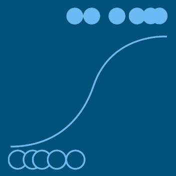
All published experiments refer to a subset of 14 of these attributes (13 features and 1 target variable), which are available via the UCI machine learning database

Data science experiments have concentrated on distinguishing the presence and absence of heart disease based on the 13 features

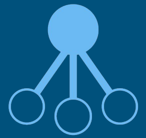
Feature	Variable	Description	Type
Age	age	In years	Continuous
Sex	sex	Gender	Categorial
Chest pain type	cp	Scale of 0 to 4 (typical angina, atypical angina, non-angina pain, asymptomatic)	Categorial
Resting blood pressure	trestbps	Diastolic blood pressure in mmHg	Continuous
Cholesterol	chol	Serum cholesterol (mg/dl)	Continuous
Fasting blood sugar	fbs	Greater than 120mg/dl, value of 0 or 1	Categorical
Resting ECG	restecg	Value of 0, 1, or 2	Categorical
Maximum heartrate achieved	thalach	Maximum heartrate from thallium test[i]	Continuous
Exercise-induced angina	exang	Value of 0 or 1	Categorical
Old-peak	oldpeak	ST depression induced by exercise relative to rest	Continuous
Slope-peak	slope	Slope of peak exercise ST segment, value of 1, 2, or 3	Categorical
Coronary artery disease	ca	Number of major vessels (0-3) colored by fluoroscopy	Categorical
Exercise thallium	thal	Exercise thallium scintigraphic defects, vales of 3 (normal), 6 (fixed defect), or 7 (reversible defect)	Categorial

Methodological approach

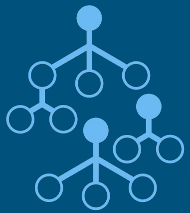
Emulate classification models that have shown the most promising performance in other studies



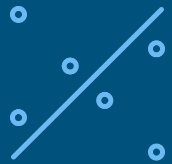
Logistic Regression



Naive Bayes



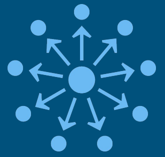
Random Forest



Support Vector Machines

Attempt to improve accuracy by synthesizing more cases on original distribution and tuning parameters

Parameter Tuning



Synthetic Data

Summary statistics

	n	min	mean	median	max	sd
age	6060	29	54.327723	55.0	77.0	8.967815
trestbps	6060	94	131.179043	130.0	200.0	17.335760
chol	6060	126	247.889604	244.0	564.0	53.500861
thalach	6060	71	149.407591	153.0	202.0	23.180276
oldpeak	6060	0	1.052541	0.8	6.2	1.146270

cp	ca	restecg	slope	thal
0:2886	0:3447	0:3052	0: 501	0: 42
1: 966	1:1370	1:2913	1:2864	1: 429
2:1720	2: 742	2: 95	2:2695	2:3267
3: 488	3: 400	NA	NA	3:2322
NA	4: 101	NA	NA	NA

exang	fbs	sex	target
0:4071	0:5165	0:1895	0:2907
1:1989	1: 895	1:4165	1:3153

- Based on the distributions of n = 303 observations in the original dataset, n = 6,060 cases were simulated in the synthetic dataset
- No missing data or NAs
- As expected, both original and synthetic datasets have similar shape and summary statistics (mean, sd, min, max)

Logistic Regression Model

Background:

- Regression technique to assign observations to a discrete set of categories based on predictor variables
- Contribution of individual predictors to overall fit can be interpreted

Approach and findings:

- Prepared 14 models using only factor, numeric, or selected variables, training and testing on original and synthetic data
- Factorized model most accurate model for both datasets

\\ Metrics Model	Accuracy	F1	Sensitivity	Specificity	Precision
Original Data	0.813	0.781	0.735	0.878	0.833
Synthesized Data	0.793	0.767	0.717	0.862	0.826

Random Forest Model

Background:

- Generates multiple decision trees based on bootstrap sampling
- Subsampling reduces variance and random feature selection decorrelates, improving predictive accuracy and helping to control over-fitting

Approach and findings:

- Baseline model created with original data achieved accuracy of 0.796
- Used hyper-parameter tuning and cross-validation to improve performance

\\ Metrics Model	Accuracy	F1	Sensitivity	Specificity	Precision
Original Data	0.951	0.945	0.935	0.964	0.956
Synthesized Data	0.858	0.841	0.826	0.885	0.857

Support Vector Machines Model

Background:

- Supervised learning technique that defines a margin-maximizing hyperplane as a decision boundary between classes
- Works well in high dimensions, but prone to overfitting, computationally intensive, and hard to interpret

Approach and findings:

- Experimented with radial (RBF) and linear kernels
- Tuning sigma and C parameters did not augment performance
- Achieved best performance with RBF kernel on synthetic data

Model \\ Metrics	Accuracy	F1	Sensitivity	Specificity	Precision
Original Data	0.813	0.788	0.765	0.836	0.813
Synthesized Data	0.840	0.809	0.735	0.927	0.825

Naive Bayes Model

Background:

- Considers all variables to independently contribute to the probability of heart disease
- Requires small amount of training data to estimate parameters, low CPU and memory consumption

Approach and findings:

- Numeric variables `age` and `sex` removed, `chol` converted to categorical variable for improved classification
- Performance did not improve on larger synthetic dataset

\\\\ Metrics Model	Accuracy	F1	Sensitivity	Specificity	Precision
Original Data	0.787	0.742	0.677	0.878	0.821
Synthesized Data	0.786	0.765	0.751	0.815	0.779

Conclusions and summary

We saw highest accuracy from random forest model on the original dataset (n = 303), which improved markedly based on hyperparameter tuning

SVM exceeded the median accuracy of the study pool, but our Naive Bayes and Logistic Regression implementations did not

The synthesized data did not universally lead to higher accuracy or stability of models - SVM was the only model that improved

Technique	Median Accuracy (n = 62 studies)	Highest Accuracy (n = 62 studies)	Our Best Accuracy
Logistic Regression	0.855	0.855	0.813 (original)
Random Forest	0.724	0.814	0.951 (original)
Support Vector Machine	0.809	0.875	0.840 (synthetic)
Naive Bayes	0.819	0.950	0.787 (original)